

Проведение анализа временного ряда

1 Постановка задачи

Необходимо:

1. Проверить является ли временной ряд стационарным в широком смысле, используя следующие методы:
 - (a) Визуальная оценка, согласно которой нужно отрисовать ряд и скользящую статистику (среднее, стандартное отклонение). А также построить график, на котором будет отображен сам ряд и различные скользящие статистики.
 - (b) Тест Дики - Фуллера.
2. Разложить временной ряд на тренд, сезонность, остаток в соответствии с аддитивной, мультипликативной моделями. Визуализировать их, оценить стационарность получившихся рядов, сделать выводы.
3. Проверить является ли временной ряд интегрированным порядка k . Если является, применить к нему модель ARIMA, подобрав необходимые параметры с помощью функции автокорреляции и функции частичной автокорреляции. Выбор параметров обосновать. Отобрать несколько моделей. Предсказать значения для тестовой выборки. Визуализировать их, посчитать r^2 score для каждой из моделей. Произвести отбор наилучшей модели с помощью информационного критерия Акаике. Провести анализ получившихся результатов.

Будем называть **временным рядом** совокупность значений какого-либо показателя за несколько последовательных периодов времени. Отличительная особенность статистического анализа временных рядов состоит в том, что последовательность наблюдений $y(t_1), y(t_2), \dots, y(t_r)$ рассматривается как реализация последовательности, вообще говоря, статистически зависимых случайных величин. Чтобы сделать задачу статистического анализа временных рядов доступной для практического решения, приходится ограничивать класс рассматриваемых моделей временных рядов, вводя те или иные предположения относительно структуры ряда и структуры его вероятностных характеристик. Одно из таких ограничений предполагает стационарность временного ряда. Под стационарностью мы будем понимать, что у временного ряда некоторые свойства не меняются с течением времени. В соответствии с этим рассмотрим 2 типа стационарности:

1. **Строгая стационарность**, или стационарность в узком смысле.

Будем называть случайный процесс строго стационарным, если сдвиг во времени не меняет ни одну из функций плотности распределения. То есть, если ко всем моментам времени добавить некоторую целочисленную величину, то сама функция плотности не изменится, $f_n(x_{t_1}, \dots, x_{t_n}) = f_n(x_{t_1+\delta}, \dots, x_{t_n+\delta})$ для всех n , моментов времени t_1, \dots, t_n и целочисленных δ . Если плотности распределения не существует, можно сформулировать это определение в терминах функции распределения. Поскольку при каждом фиксированном t случайный процесс есть случайная величина, то для каждой из них можно рассматривать такие характеристики, как мат. ожидание, дисперсию, автоковариационную и автокорреляционную функции. **Автоковариационной функцией** случайного

процесса называется совокупность значений ковариаций при всевозможных значениях расстояния между моментами времени. Если случайный процесс стационарен в узком смысле, то у него:

- математическое ожидание не зависит от времени;
- дисперсия не зависит от времени;
- автоковариационная и автокорреляционные функции: а) зависят только от сдвига, от разности моментов времени; б) являются четными функциями.

Коэффициент корреляции $\rho(\tau) = \frac{\gamma(\tau)}{\gamma(0)}$ между разделёнными на τ значениями временного ряда, где функцию $\gamma(\tau)$ можно рассматривать как всевозможные значения автоковариаций, где τ пробегает целочисленные значения от $-\infty$ до $+\infty$, а $\gamma(0) = Cov(X_{t_1}, X_{t_1}) = \sigma^2$ - это ковариация, разделенная на корень из произведения двух дисперсий, но так как дисперсия постоянна, то мы получаем просто σ^2 . Это выражение определяет **автокорреляционную функцию** временного ряда.

2. Слабая стационарность, или стационарность в широком смысле.

Если случайный процесс таков, что у него математическое ожидание и дисперсия существуют и не зависят от времени, а автокорреляционная (автоковариационная) функция зависит только от разности значений $(t_1 - t_2)$, то такой процесс мы назовем стационарным в широком смысле, или слабо стационарным.

Следовательно, построив графики ряда и скользящей статистики, мы сможем определить является ли ряд стационарным в широком смысле. Также это можно сделать с помощью теста Дики-Фуллера. Для этого введём следующие понятия.

Тренд — тенденция изменения показателей временного ряда. Тренды могут быть описаны различными функциями — линейными, степенными, экспоненциальными и т. д. Тип тренда устанавливают на основе данных временного ряда, путем осреднения показателей динамики ряда, на основе статистической проверки гипотезы о постоянстве параметров графика.

Авторегрессионная (AR-) модель - модель временных рядов, в которой значения временного ряда в данный момент линейно зависят от предыдущих значений этого же ряда. Авторегрессионный процесс порядка p (AR(p)-процесс) определяется следующим образом

$$X_t = c + \sum_{i=1}^p a_i X_{t-i} + \varepsilon_t,$$

где a_1, \dots, a_p — параметры модели (коэффициенты авторегрессии), c — постоянная (часто для упрощения предполагается равной нулю), а ε_t — белый шум. Процессом **белого шума** называют стационарный случайный процесс X_t , $t = 0, \pm 1, \pm 2 \dots$, для которого верно:

$$(X_t) = 0$$

$$D(X_t) = \sigma^2$$

и при $t \neq 0$

$$\rho(\tau) = 0$$

Временной ряд имеет **единичный корень**, или порядок интеграции один, если его первые разности образуют стационарный ряд. Это условие записывается как $y_t \sim I(1)$ если ряд

первых разностей $\Delta y_t = y_t - y_{t-1}$ является стационарным $\Delta y_t \sim I(0)$. При помощи этого теста проверяют значение коэффициента a в авторегрессионном уравнении первого порядка AR(1)

$$y_t = a \cdot y_{t-1} + \varepsilon_t,$$

где y_t - временной ряд, а ε - белый шум. Если $a = 1$, то процесс имеет единичный корень, тогда ряд y_t не стационарен, является **интегрированным временным рядом** первого порядка - $I(1)$. Если $|a| < 1$, то ряд стационарный - $I(0)$. Приведенное авторегрессионное уравнение AR(1) можно переписать в виде:

$$\Delta y_t = b \cdot y_{t-1} + \varepsilon_t,$$

где $b = a - 1$, а Δ — оператор разности первого порядка $\Delta y_t = y_t - y_{t-1}$. Поэтому проверка гипотезы о единичном корне в данном представлении означает проверку нулевой гипотезы о равенстве нулю коэффициента b (существует единичный корень, ряд нестационарный). **Статистика теста (DF-статистика)** — это обычная t -статистика для проверки значимости коэффициентов линейной регрессии. Однако, распределение данной статистики отличается от классического распределения t -статистики. Распределение DF-статистики называется **распределением Дики — Фуллера**.

Существует три версии теста (тестовых регрессий):

1. Без константы и тренда: $\Delta y_t = b \cdot y_{t-1} + \varepsilon_t$
2. С константой, но без тренда: $\Delta y_t = b_0 + b \cdot y_{t-1} + \varepsilon_t$
3. С константой и линейным трендом: $\Delta y_t = b_0 + b_1 \cdot t + b \cdot y_{t-1} + \varepsilon_t$

Для каждой из трёх тестовых регрессий существуют свои критические значения DF-статистики, которые берутся из специальной таблицы Дики — Фуллера (МакКиннона). Если значение статистики лежит левее критического значения (критические значения — отрицательные) при данном уровне значимости, то нулевая гипотеза о единичном корне отклоняется и процесс признается стационарным (в смысле данного теста). В противном случае гипотеза не отвергается и процесс может содержать единичные корни, то есть быть нестационарным (интегрированным) временным рядом.

Модель с распределённым лагом — это модель временного ряда, в которой в уравнение регрессии включено как текущее значение объясняющей переменной, так и значения этой переменной в предыдущих периодах.

Если в тестовые регрессии добавить лаги первых разностей временного ряда, то распределение DF-статистики (а значит, критические значения) не изменится. Такой тест называют **расширенным тестом Дики — Фуллера (ADF)**. Необходимость включения лагов первых разностей связана с тем, что процесс может быть авторегрессией не первого, а более высокого порядка. Рассмотрим на примере модели AR(2):

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + \varepsilon_t.$$

Данную модель можно представить в виде:

$$\Delta y_t = (a_1 + a_2 - 1) y_{t-1} - a_2 \Delta y_{t-1} + \varepsilon_t.$$

Если временной ряд имеет один единичный корень, то первые разности по определению стационарны. А поскольку y_{t-1} по предположению нестационарен, то если коэффициент при

нём не равен нулю, уравнение противоречиво. Таким образом, из предположения об интегрированности первого порядка для такого ряда следует, что $a_1 + a_2 - 1 = 0$. Таким образом, для проверки наличия единичных корней в данной модели следует провести стандартный DF-тест для коэффициента при y_{t-1} , причем в тестовую регрессию должен быть добавлен лаг первой разности зависимой переменной.

Приступим к разложению ряда на составляющую тренда, сезонную компоненту и оставшуюся нерегулярную составляющую. Этот процесс называется **сезонной декомпозицией**.

Общая модель. Функцию исходного ряда можно разложить на следующие компоненты:

- $x(t)$ – **тренд**, устойчивая долговременная тенденция изменения значений временного ряда, закономерно изменяющаяся во времени;
- $s(t)$ – **сезонная** составляющая, периодически повторяющаяся компонента временного ряда, на которую влияют погодные условия, социальные привычки, религиозные традиции и прочее;
- $z(t)$ – **остаток** – величина, показывающая нерегулярную (не описываемую трендом или сезонностью) составляющую исходного ряда в определённом временном интервале.

Конкретные функциональные взаимосвязи между этими компонентами могут иметь самый разный вид. Однако можно выделить два основных способа, с помощью которых они могут взаимодействовать:

Аддитивная модель: $y(t) = x(t) + s(t) + z(t)$

Мультипликативная модель: $y(t) = x(t) * s(t) * z(t)$

Здесь $y(t)$ обозначает значение временного ряда в момент времени t .

Процесс построения модели включает в себя следующие шаги:

1. Выравнивание исходного ряда **методом скользящей средней**.
2. Расчет значений сезонной компоненты $s(t)$.
3. Устранение сезонной компоненты из исходных уровней ряда и получение выровненных данных $y(t) - s(t) = x(t) + z(t)$ в аддитивной или $\frac{y(t)}{s(t)} = x(t) * z(t)$ в мультипликативной модели.
4. Аналитическое выравнивание уровней $x(t) + z(t)$ или $x(t) * z(t)$ и расчет значений $x(t)$ с использованием полученного уравнения тренда.
5. Расчет полученных по модели значений $x(t) + z(t)$ или $x(t) * z(t)$.
6. Расчет абсолютных и/или относительных ошибок. Если из временного ряда удалить тренд и сезонную составляющую, то останется нерегулярная компонента, так называемая, ошибка. Если полученные значения ошибок не содержат автокорреляции, ими можно заменить исходные уровни ряда и в дальнейшем использовать временной ряд ошибок для анализа взаимосвязи исходного ряда и других временных рядов.

Модель скользящего среднего q -го порядка $MA(q)$ — модель временного ряда вида:

$$X_t = \sum_{j=0}^q b_j \varepsilon_{t-j},$$

где ε_t — белый шум, b_j — параметры модели (b_0 можно считать равным 1 без ограничения общности).

Также в модель иногда добавляют константу. Тем не менее, поскольку чаще всего модели скользящего среднего используются для моделирования случайных ошибок временных рядов, то константу можно считать параметром основной модели.

Процесс белого шума формально можно считать процессом скользящего среднего нулевого порядка — $MA(0)$.

Чаще всего на практике используют процесс скользящего среднего первого порядка $MA(1)$:

$$X_t = \varepsilon_t + b\varepsilon_{t-1}$$

Метод скользящих средних

Прежде, чем рассчитывать сезонную компоненту, исходный временной ряд необходимо выровнять методом скользящих средних;

Вычисляя скользящее среднее для временного ряда, **интервал сглаживания** (ширина окна) берется равным периоду сезонности. Если период сезонности — четное число, можно в случае простого скользящего среднего, провести процедуру центрирования, которая заключается в повторном скользящем с шагом, равным двум. Число уровней сглаженного ряда будет меньше на величину шага скользящей средней.

После определения скользящих средних вся сезонная (т.е. внутри сезона) изменчивость будет исключена и поэтому разность (в случае аддитивной модели) или отношение (для мультипликативной модели) между наблюдаемым Y_i и сглаженным рядом \hat{Y}_t будет выделять сезонную составляющую плюс нерегулярную компоненту.

Таким образом, результатом процедуры сглаживания будет временной ряд выровненных значений \hat{Y}_t , не содержащий сезонной компоненты. То есть: ряд скользящих средних вычитается из наблюдаемого ряда (в аддитивной модели) или же значения наблюдаемого ряда делятся на значения скользящих средних (в мультипликативной модели).

Модель ARIMA

Если ряд после взятия d последовательных разностей приводится к стационарному, то назовем этот ряд $ARIMA(p, d, q)$. ARIMA - процесс авторегрессии - интегрированного скользящего среднего. При этом p - параметр AR-части, d - степень интеграции, и q - это параметр MA-части. В операторном виде ARIMA (p, d, q) записывается как:

$$\alpha_p(L)\Delta^d x_t = \beta_q(L)\varepsilon_t$$

Этот процесс нестационарный, потому что здесь не выполняется условие, что все корни характеристического уравнения по модулю меньше единицы. Но, если обозначить $(1 - L)^d x_t = y_t$, то y_t - стационарный процесс. Задачу построения модели типа ARIMA по реализации случайного процесса можно разбить на несколько этапов.

I этап

1. Установить порядок интеграции d , то есть добиться стационарности ряда, взяв достаточное количество последовательных разностей.
2. После этого мы получаем временной ряд Y к которому нужно подобрать уже ARMA(p, q). Исходя из поведения автокорреляционной и частной автокорреляционной функций, установить параметры p и q .

I этап принято называть идентификацией модели $ARIMA(p,d,q)$. Это всего лишь определение величин p , d , q , но именно в такой последовательности: сначала d , а потом p и q .

II этап Оценка коэффициентов $\alpha_1, \alpha_2, \dots, \alpha_p, \beta_1, \beta_2, \dots, \beta_q$ при условии, что мы уже знаем p и q .

III этап По остаткам осуществляется тестирование или диагностика построенной модели.

IV этап Использование модели, в основном, для прогнозирования будущих значений временного ряда. Если исследуемый ряд нестационарный, то его автокорреляционная функция не будет убывать. Если ряд стационарен, то мы знаем, что, начиная с какого-то номера, теоретические автокорреляции будут убывать. Поэтому можно рассчитать их оценки - выборочные автокорреляции, посмотреть, убывают они или нет. Если ряд окажется стационарным, перейти к определению параметров p и q . Если нет, то надо построить ряд первых разностей и проверить на стационарность его. Сначала выбираем значение порядков дифференцирования d так, чтобы ряд стал стационарным. Выбор параметра q проводим с помощью графика функции автокорреляции. q - номер последнего несезонного лага, при котором автокорреляция значима. Параметр p выбираем по графику функции частичной автокорреляции. **Частичная автокорреляция** - автокорреляция после снятия регрессии предыдущего порядка. p задается как номер последнего несезонного лага, при котором частичная автокорреляция значима. Далее будем перебирать разные модели в классе $ARIMA$ с выбранным значением параметра d , начиная с тех начальных приближений, которые мы получили из автокорреляционных функций. Сравнивать разные модели будем по **информационному критерию Акаике (AIC)**:

$$AIC = 2k - 2\ln(L),$$

где k — число параметров в статистической модели, и L — максимизированное значение функции правдоподобия модели. Оптимальной по критерию Акаике будет модель, у которой значение этого критерия самое маленькое из всех возможных, потому что такая модель будет достаточно хорошо описывать данные, с одной стороны, и содержать не слишком большое количество параметров, с другой.

2 Программный подход к решению задачи

Первым шагом мы считываем данные из таблиц `training.xlsx`, `testing.xlsx` с помощью метода `read_excel()` из библиотеки `Pandas`.

Используем метод `Value.rolling().mean()` для получения среднего значения временного ряда и метод `Value.rolling.std()` для скользящего отклонения. Составляем графики найденных величин с помощью библиотеки `matplotlib`. Затем проводим расширенный тест Дики-Фуллера, пользуясь методом `adfuller` из библиотеки `statsmodels`, версия без константы и тренда. Находим, что у ряда есть единичные корни по значениям p -value и `adf`, значит он нестационарен.

С помощью функции из библиотеки `statsmodels` - `seasonal_decompose` с параметрами: значения из таблицы, типами мультипликативный и аддитивный и без фильтра - раскладываем ряд на тренд и сезональность.

С помощью метода `diff` из исходного ряда получаем ряд разностей `datadiff`, а с помощью метода `dropna` удаляем из него все нулевые элементы.

Проводим тест Дики-Фуллера для `datadiff`, получаем, что он стационарен, следовательно, исходный ряд - интегрированный ряд первого порядка. Рисуем `datadiff`. Замечаем, что тренд отсутствует, это подтверждает, что ряд первых разностей является стационарным, а наш

исходный ряд — интегрированным рядом первого порядка. С помощью функции `plot_acf` из библиотеки `statsmodels` рисуем функции автокорреляции и частичной автокорреляции. Используя функцию `ARIMA` из библиотеки `statsmodels`, создаём `model`.

Первый параметр функции `ARIMA` - значения таблицы, а второй-тройка параметров p, d, q . Параметры p и q находим с помощью графиков функции частичной и полной автокорреляции. По коррелограмме `ACF` определяем q = количество автокорреляционных коэффициентов сильно отличных от нуля в модели `MA` - оно равно 2. По коррелограмме `PACF` можно определить p = максимальный номер коэффициента, сильно отличного от нуля в модели `AR` - оно равно 2. Получаем характеристику модели `model` с помощью метода `summary()`. Используем функции `plot_acf`, `autocorrelation_plot` нарисуем лаги ряда и автокорреляцию. Выберем часть таблицы, начиная с 1986-08-11 и спрогнозируем её с помощью метода `predict(model)`. Функция `r2_score` библиотеки `sclearn` возвращает параметр r^2 , характеризующий точность предсказаний.

Далее используем метод `predict` и получаем предсказание для продолжения таблицы, нарисуем его. Теперь точно так же сделаем другое предсказание, только возьмем параметры $p = 1, q = 1$ для части таблицы, начиная с 1988-03-01. Найдём для него `r2_score` и визуализируем предсказание, получим характеристику модели `ARIMA`. Сравним параметры `AIC` двух предсказаний.

3 Выводы

Для данного нам временного ряда мы построили графики скользящей статистики и по ним определили, что ряд нестационарен, так как его математическое ожидание и дисперсия зависят от времени. Нестационарность данного ряда подтверждает и тест Дики-Фуллера. Значение статистики лежит правее критического значения при уровне значимости 5%, поэтому нулевая гипотеза о единичном корне не отклоняется, то есть процесс может содержать единичные корни. Достоверность статистики больше уровня значимости.

После разложения временных рядов на тренд, сезонность и остаток, мы заметили, что для обеих моделей оригинальный ряд и трендовая составляющая нестационарны, так как их скользящая статистика зависит от времени. Однако так же для обеих моделей справедлива стационарность сезонной составляющей и остатка, их скользящая статистика не зависит от времени.

Наилучшее значение $r^2score = 1$. Если оно будет отклоняться вправо или влево, то велика вероятность неточности построения модели. Более удачной моделью будет модель с меньшим `AIC`. В нашем случае $AIC_2 = 248 < AIC_2 = 252$, следовательно вторая модель предпочтительнее.

4 Работу выполнили

1. Алеся Евенко - написание `README`
2. Михаил Таипов - написание кода, помощь в создании `README`
3. Тимур Насыров - помощь в написании кода